MULTIMEDIA UNIVERSITY

STUDENT ID NO

MULTIMEDIA UNIVERSITY

FINAL EXAMINATION

TRIMESTER 2, 2016/2017

TNL3221 – NATURAL LANGUAGE PROCESSING

(All Sections / Groups)

2 MARCH 2017
2.30 p.m. – 4.30 p.m.
(2 Hours)

INSTRUCTIONS TO STUDENTS

1. This question paper consists of 5 pages with 4 questions only.
2. Attempt ALL questions. All questions carry equal marks and the distribution of the marks for each question is given.
3. Please write all your answers in the Answer Booklet provided.

Question 1

(a) Briefly describe FOUR applications of natural language processing. [4 marks]

(b) What distinguishes language processing applications from other data processing systems? Give an example. [2 marks]

(c) How many token, type, bigram, and trigram (including punctuation) are there in the following sentence?

> She said she resorted to escaping as she was in great fear and did not have time to remember the registration number of the lorry.

[2 marks]

(d) Parts-of-speech can be divided into two broad super categories: closed class types and open class types. Briefly explain these two class types. [2 marks]

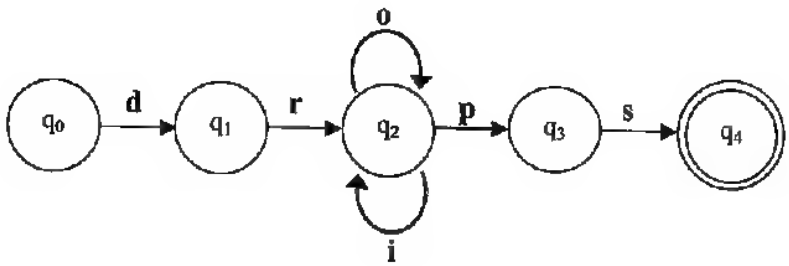
Question 2

(a) Write a regular expression for each of the following: [4 marks]

i. That matches the strings: Hello Kitty, Hello kitty, hello Kitty, or hello kitty.
ii. That finds lines consisting only of letters a, b or c.
iii. That matches any character between *'s'* and *'ng'*.
iv. That matches decimal numbers (examples: 123.45, 987.6, 0.015).

(b) Based on the following automaton:

i. Draw a state-transition table. [4 marks]
ii. Write the regular expression. [1 mark]
iii. Justify whether it is a deterministic finite state automaton (DFSA) or non-deterministic finite state automaton (NFSA). [1 mark]



Continued...

Question 3

(a) The following table shows the frequencies of frequencies c, N_c where $c = 0, 1,\ldots,10$. Fill in the blanks by calculating the smoothed count c^*(GT) based on Good-Turing estimates. Show the calculation steps. [5 marks]

(Note: Students only need to copy the column c^* (GT) in your answer booklet.)

c (MLE)	N_c	c^* (GT)
0	74,513,701	
1	37,365	
2	5,820	
3	2,111	
4	1,067	
5	719	
6	468	
7	330	
8	250	
9	179	
10	160	-

Table 1: Frequencies of frequencies

(b) Given the following context-free grammar:

$$S \rightarrow NP\ VP$$

$$NP \rightarrow Pronoun \mid Proper\text{-}Noun \mid Det\ Nominal$$

$$Nominal \rightarrow Nominal\ Noun \mid Noun$$

$$VP \rightarrow Verb \mid Verb\ NP \mid Verb\ NP\ PP \mid Verb\ PP$$

$$PP \rightarrow Preposition\ NP$$

Use the shift-reduce approach to search this string: '*that product arrived on Monday*' [5 marks]

Continued...

Question 4

(a) Use Penn Treebank tagset to tag each word of this sentence: [6 marks]
It took a man with extraordinary qualities to succeed in Malaysia.

1.	CC	Coordinating conjunction	25.	TO	*to*
2.	CD	Cardinal number	26.	UH	Interjection
3.	DT	Determiner	27.	VB	Verb, base form
4.	EX	Existential *there*	28.	VBD	Verb, past tense
5.	FW	Foreign word	29.	VBG	V, gerund/pres. particle
6.	IN	Preposition/sub. conjunction	30.	VBN	V, past particle
7.	JJ	Adjective	31.	VBP	V, non-3rd ps. sing. pres.
8.	JJR	Adjective, comparative	32.	VBZ	V, 3rd-ps. sing. present
9.	JJS	Adjective, superlative	33.	WDT	*wh*-determiner
10.	LS	List item marker	34.	WP	*wh*-pronoun
11.	MD	Modal	35.	WP$	Possessive *wh*-pronoun
12.	NN	Noun, singular or mass	36.	WRB	*wh*-adverb
13.	NNS	Noun, plural	37.	#	Pound sign
14.	NNP	Proper noun, singular	38.	$	Dollar sign
15.	NNPS	Proper noun, plural	39.	.	Sentence final punct.
16.	PDT	Predeterminer	40.	,	Comma
17.	POS	Possessive ending	41.	:	Colon, semicolon
18.	PRP	Personal pronoun	42.	(	Left bracket character
19.	PPS	Possessive pronoun	43.	)	Right bracket character
20.	RB	Adverb	44.	"	Straight double quote
21.	RBR	Adverb, comparative	45.	‘	Left open single quote
22.	RBS	Adverb, superlative	46.	“	Left open double quote
23.	RP	Particle	47.	’	Right open single quote
24.	SYM	Symbol	48.	”	Right open double quote

Table 2: Penn TreeBank Tagset

Continued...

(b) Given the following unigram and bigram counts:

I	Want	To	Eat	Chinese	Food	Lunch	Thai
3437	1215	3256	938	213	1506	459	315

Table 3: Unigram Counts

	I	Want	To	Eat	Chinese	Food	Lunch	Thai
I	8	1087	0	13	0	0	0	0
Want	3	0	786	0	6	8	6	8
To	3	0	10	860	3	0	12	5
Eat	0	0	2	0	19	2	52	28
Chinese	2	0	0	0	0	120	1	0
Food	19	0	17	0	0	0	0	0
Lunch	4	0	0	0	0	1	0	0
Thai	3	0	0	0	0	150	2	0

Table 4: Bigram Counts

i. Compute the bigram probabilities P(Want|I), P(To|Want), P(Eat|To), P(Thai|Eat), and P(Food|Thai). [Round your answers to two decimal places.] [2.5 marks]

ii. Suppose P(I|<s>) = 0.25 and P(</s>|Food) = 0.68, calculate the probability of the sentence "I Want To Eat Thai Food". [Round your final answer to five decimal places.] [1.5 marks]

End of Paper